Investigating harbor porpoise (*Phocoena phocoena*) population differentiation using RAD-tag genotyping by sequencing

Ljerka Lah[1], Harald Benke[2], Per Berggren[3], Þorvaldur Gunnlaugsson[4], Santiago Lens[5], Christina Lockyer[6], Ayaka Amaha Öztürk[7], Bayram Öztürk[7], Iwona Pawliczka[8], Anna Roos[9], Ursula Siebert[10], Krzysztof Skóra[8], Ralph Tiedemann[1]

[1]Unit of Evolutionary Biology/Systematic Zoology, Institute of Biochemistry and Biology, University of Potsdam, D-14476 Potsdam, Germany

[2]Deutsches Meeresmuseum, 18439 Stralsund, Germany

[3]School of Marine Science and Technology, Newcastle University, Newcastle NE1 7RU, UK

[4]Marine Research Institute, IS-101 Reykjavík, Iceland

[5]Instituto Espanol de Oceanografia, Centro Oceanográfico de Vigo, E-36390 Vigo, Spain

[6]North Atlantic Marine Mammal Commission, N-9294 Tromsø, Norway

[7]Marine Biology Department, Faculty of Fisheries, Istanbul University, Istanbul, Turkey

[8]Hel Marine Station, University of Gdansk, 84-150 Hel, Poland

[9]Swedish Museum of Natural History, S-104 05 Stockholm, Sweden

[10]Institute for Terrestrial and Aquatic Wildlife Research (ITAW), University of Veterinary Medicine Hannover, 25761 Büsum, Germany

ABSTRACT

The population status of the harbor porpoise (*Phocoena phocoena*) in the Baltic Sea and adjacent regions is still not fully resolved. Here, we present a pilot study using the double digest restriction-site associated DNA sequencing (ddRAD-seq) genotyping-by-sequencing method on specimens from the Baltic Sea, eastern North Sea, Spain and the Black Sea. From a single Illumina lane and a set of 49 individuals, we obtained around 6000 SNPs. We used these markers to estimate population structure and differentiation, and identified splits between porpoises from the North Sea and the Baltic, and within regions in the Baltic Sea (between the Belt Sea and the Inner Baltic Sea). The SNP analysis confirms population structure elucidated by previous mtDNA/microsatellite studies. We demonstrate the feasibility of SNP analysis on opportunistically sampled cetacean samples, with varying DNA quality, for population diversity and divergence analysis.

INTRODUCTION

The harbor porpoise (*Phocoena phocoena*) population structure in the Baltic Sea relative to the adjacent western Skagerrak (SKA) region and the North Sea (NOS) population has been a continuous matter of debate particularly with regard to conservation management practices. While the eastern North Sea population behaves as a continuous population with significant isolation-by-distance (Fontaine *et al*, 2007), strong barriers to gene flow may exist in the Baltic Sea and adjacent regions (Wiemann *et al*, 2009). These regions are a series of relatively deep basins: the Kattegat (KAT), the Belt Sea (BES) and the Inner Baltic Sea (IBS), separated by shallower underwater ridges which could be oceanographic barriers hindering gene flow (Fig. 1).

Porpoise population differentiation between these regions, based on morphology or genetics, has been inferred in several preceding studies (Wiemann *et al*, 2009 and references therein). Wiemann *et al.* (2009) have conducted one of the most recent and comprehensive ones, using mtDNA haplotypes and 15 microsatellite loci from nearly 500 and 305 individuals, respectively. They showed that the NOS and Baltic Sea populations were not panmictic. Further, they identified a population split between the SKA and BES regions, and a possible further split between the BES and the IBS regions. Here, we revisit this study by using a subset of the same samples along with additional samples from Turkey, Spain and Iceland to perform restriction-site associated DNA sequencing (RAD-seq)-based population genomic analyses.

RAD-seq has become one of the most widely used genotyping methods in population genomics studies of non-model organisms (Davey *et al*, 2011). It combines reduced representation library construction, achieved through restriction enzyme (RE) digestion of genomic DNA at conserved sites, and Next Generation Sequencing (NGS) methods. Traditional RAD-seq uses a single RE digest coupled with secondary random fragmentation to generate NGS libraries for single-end or paired-end sequencing (Baird *et al*, 2008; Etter *et al*, 2011). Double digest RAD sequencing (ddRAD-seq), uses a two enzyme double digest followed by a more precise size selection step which allows greater control of the fraction of regions represented in the final library and ensures better reproducibility (Peterson *et al*, 2012).

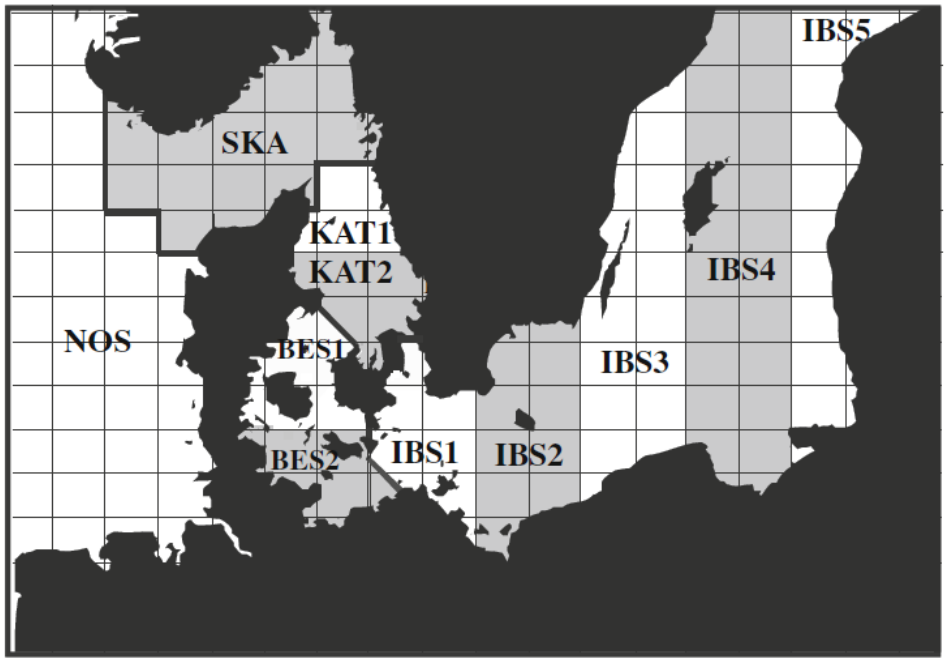

Figure 1. Sampling locations (50 km x 50 km grids defined by the International Council for the Exploration of the Sea, ICES) and assignment to regions (solid lines). Regions are North Sea (NOS), Skagerrak (SKA), Kattegat (KAT), Belt Sea (BES), and Inner Baltic Sea (IBS). Except for the distinction between NOS and SKA, all boundaries (solid lines) are defined by shallow underwater ridges (up to 50 m depth) between SKA and KAT, KAT and BES, and the Darss sill between BES and IBS. Within KAT, BES, and IBS, regions are further divided into sub-regions of 100 km width (subsequently numbered and indicated by alternating white/gray coloring). Reproduced from Wiemann *et al* (2009).

In this study, we used a modified ddRAD-seq approach (Sonah *et al*, 2013) combined with the STACKS bioinformatics analysis pipeline (Catchen *et al*, 2011), which was specifically designed to deal with a wide array of RAD-tag sequencing data. The main objective of our study was to evaluate whether the RAD-tag genotyping-by-sequencing method is appropriate for the study of population differentiation among the harbor porpoise populations of the Baltic Sea and adjacent regions by attempting to reproduce (or improve) the results of previous studies. Furthermore, given that the sampling of porpoises is necessarily opportunistic, since samples originate from by-caught or stranded individuals in various states of decomposition, we aimed to establish what effect DNA degradation has on the sequencing output and final number of usable SNP loci for population genomics analyses.

METHODS

Sample selection and DNA extraction

We analyzed 49 samples from either by-caught or stranded individuals. Most of these samples were from the Baltic Sea and adjacent regions (North Sea – NOS, Skagerrak – SKA, Kattegat – KAT, Belt Sea – BES and the Inner Baltic Sea – IBS), of which the majority has been already used in a previous study (Wiemann *et al*, 2009). The rest were from Turkey, Spain, and Iceland (see Appendix for detailed sample information).

We extracted total genomic DNA from approximately 25 mg of tissue from samples stored at -20 ^{0}C (frozen or stored in ethanol) using the NucleoSpin Tissue Kit (Macherey-Nagel, Germany) following the manufacturer's recommendations. We measured DNA concentration using a NanoDrop 1000 (Thermo Scientific, USA). Using the Agilent 2200 TapeStation with the Genomic ScreenTape System (Agilent Technologies, USA), we additionally assessed sample quality and quantity.

Genotyping by sequencing and data analyses

The ddRAD-tag libraries were prepared by a commercial sequencing service provider (LGC Genomics, Berlin) using total genomic DNA and the restriction enzymes *Pst*I (rare cutter) and *Msp*I (common cutter). Briefly, the DNA samples were normalized and simultaneously digested with both enzymes. This step was followed by adapter ligation, where the *Pst*I adapter contained a unique sample barcode. The reaction mix clean-up was followed by an amplification step to add the flow cell binding sites. This step included a concurrent reduction of the amount of samples to be sequenced by elongating one of the two PCR primers by two bases (Sonah *et al*, 2013). The individual samples were then pooled and cleaned. Sequencing was preceded by a size selection step (low-melting point (LMP) agarose) to remove fragments smaller than 250 bp and larger than 500 bp.

The libraries were sequenced on one lane of the Illumina HiSeq 2000 platform (Illumina Inc., USA) with the 100 bp paired-end read module. Raw Illumina reads were processed using the CASAVA v. 1.8.2 software (Illumina Inc., USA). Samples were de-multiplexed with inline barcodes using LGC-developed software and clipped to remove Illumina TruSeq™ adapters and inline barcode remnants of all reads. Reads shorter than 20 bases were discarded and the remaining mate read stored in a separate FASTQ file for single reads. FastQC reports (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) containing read quality metrics were generated for all FASTQ files.

We processed the sequenced data and grouped the reads from all individuals using several programs from the STACKS v. 1.13 software package for analyzing RAD-seq data (Catchen *et al*, 2011, 2013). First, using the STACKS program process_radtags we filtered for read quality and trimmed the de-multiplexed paired-end and single reads to a length of 85 bp. We concatenated all three read files into one common FASTQ file per individual. With the wrapper program denovo_map.pl which can handle data without a reference genome, and which executes the three STACKS components (ustacks, cstacks and sstacks) we identified alleles in our populations set. The ustacks program aligns short sequence reads into matching stacks from which loci are formed and SNPs detected at each locus. We tested several combinations of parameter settings. For the final data analyses, the minimum depth of coverage required to create a stack and the maximum distance (in nucleotides) allowed between stacks were both set to 3; the removal of highly repetitive RAD-tags was enabled. A catalog of all loci across all individuals was created with the cstacks program with two mismatches allowed between loci when building the catalog. The sstacks program

then matched loci from each individual back to the catalog. We calculated population genetics statistics using the populations program. Loci were retained if they were present in all individuals and if the depth of coverage at each locus was equal or higher than 6 reads per individual. The populations program also enables output in several common file formats for downstream population genomics or phylogenetic analyses, such as GENEPOP and STRUCTURE formats. To compute pairwise F_{ST} values, we imported the set of SNPs into ARLEQUIN (Excoffier & Lischer, 2010).

Due to the computational limitations of handling a large number of loci in the current STRUCTURE software package v. 2.3.4 (http://pritchardlab.stanford.edu/structure.html; (Pritchard *et al*, 2000; Falush *et al*, 2003; Hubisz *et al*, 2009)), we randomly selected 3000 retained loci in 45 individuals, assigned to 8 populations (Turkey, Spain, Iceland, NOS, SKA, KAT, BES, IBS) as input (4 samples with a relatively low number of SNP loci were left out in further analysis). To streamline batch mode analyses of population structure by setting up multiple iterations for various values of parameter *K*, we used the freely available program StrAuto (www.crypticlineage.net/pages/software.html). For analyses, we ran 15 000 burn-in iterations and 150 000 MCMC repetitions, with 5 replicates for each value of *K*. *K* ranged from 2 to 10. The StrAuto output builds a zip archive containing all result files which we uploaded to STRUCTURE HARVESTER (http://taylor0.biology.ucla.edu/structureHarvester/) (Earl & VonHoldt, 2011), a program for visualizing STRUCTURE output and implementing the Evanno method (Evanno *et al*, 2005). We chose optimal values of *K* based on the Evanno deltaK result. To align multiple replicates of our data sets and facilitate the interpretation of clustering results, we used the computer program CLUMPP (CLUster Matching and Permutation Program) (Jakobsson & Rosenberg, 2007). We visualized STRUCTURE results with *distruct* v. 1.1 (http://www.stanford.edu/group/rosenberglab/distruct.html).

RESULTS

DNA quality and sequencing output

The DNA quality analyzed with Agilent TapeStation instrument revealed striking differences in DNA integrity between samples. Fragment lengths with highest intensities per sample ranged from 665 to over 19 thousand bp, with the average length of fragments for all samples being 9 923 ± 4 782 bp (standard deviation, SD).

One lane of sequencing produced over 303 million raw reads (or over 151 million raw read pairs) from 49 individuals. The average number of adapter clipped read pairs per individual was 3 023 030 ± 949 813 (SD), with the lowest numbers just above 1.7 million and the highest 5.9 million read pairs per individual. Typically, samples of low DNA quality had a lower number of read pairs. The percentage of reads removed by quality and ambiguous RAD-tags filters in process_radtags was 13.4 and 9.9, respectively, resulting in 227 171 630, or 76.7% of retained reads. Just over 12% of the reads were further removed through filters set in denovo_map.pl. Thus, 63.4% of all reads were retained for downstream processing with STACKS.

The average number of unique RAD-tag loci per individual identified by STACKS was 377 322 ± 65 872 (SD). Nearly 7% of those loci are polymorphic (Fig. 2). The average number of SNPs per individual was 33 431. After applying stringency filters in the populations program to ensure that the loci were present in all individuals from all populations with sufficient coverage, we retained a set of 6 006 loci.

Genetic diversity of harbor porpoise populations

For the loci that were polymorphic in at least one of the populations, the average major allele frequency ranged from 0.915 (Iceland) to 0.951 (Turkey) (Table 1). The respective average observed heterozygosity ranged from 0.1431 to 0.0769. The lowest levels of genetic diversity were found in the Spanish and Turkish populations which also had the lowest percentages of polymorphic loci. The highest levels were found in the Inner Baltic Sea populations (56.3%). Within the North Sea and Baltic populations, samples from SKA and KAT regions had higher levels of genetic diversity as compared to populations on either side of the transition zone between NOS and IBS.

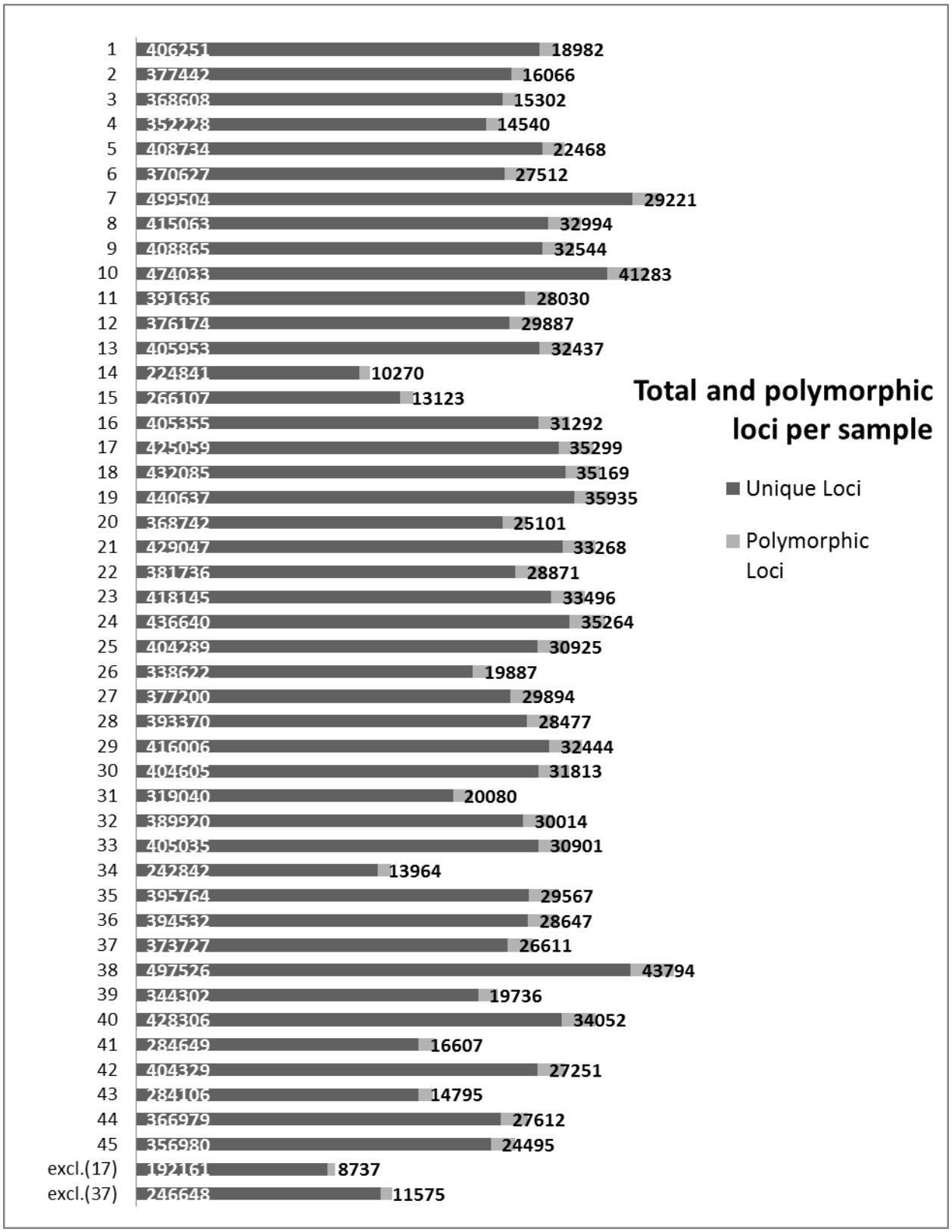

Figure 2. The number of unique and polymorphic RAD-tag loci identified with Stacks, per sample. Numbers are in order of the inclusion in the STRUCTURE analysis and refer to the Appendix.

Table 1. Summary genetics statistics calculated by the populations program for variant (polymorphic) positions (top) and all positions (bottom).

	N	Private	Nucl. Sites	% Poly. Loci	P	H_{Obs}	F_{IS}
Variant positions							
Turkey	3.9	288	1646	16.9	0.9514	0.0769	-0.0066
Spain	3.0	277	2298	23.7	0.9325	0.1161	-0.0111
Iceland	3.0	440	3190	32.8	0.9148	0.1431	-0.0038
NOS	5.9	659	4179	43.0	0.9180	0.1271	0.0088
SKA	2.0	271	2437	25.1	0.9211	0.1400	-0.0034
KAT	4.9	515	4058	41.8	0.9153	0.1363	0.0004
BES	8.9	565	4841	49.9	0.9153	0.1308	0.0023
IBS	12.7	866	5465	56.3	0.9158	0.1279	0.0075
All positions							
Turkey	4.0	288	872586	0.19	0.9995	0.0009	-0.0001
Spain	3.0	277	872585	0.26	0.9992	0.0013	-0.0001
Iceland	3.0	440	872583	0.37	0.9991	0.0016	0.0000
NOS	6.0	659	872582	0.48	0.9991	0.0014	0.0001
SKA	2.0	271	872578	0.28	0.9991	0.0016	0.0000
KAT	5.0	515	872582	0.47	0.9991	0.0015	0.0000
BES	9.0	565	872576	0.55	0.9991	0.0015	0.0000
IBS	13.0	866	872574	0.63	0.9991	0.0014	0.0001

N – average number of individuals genotyped at each locus; Private – number of variable sites unique to each population; Nucl. Sites – number of polymorphic (top) or total (bottom) nucleotide sites; % Poly. Loci – percentage of polymorphic loci; P – average frequency of the major allele; H_{Obs} – average observed heterozygosity per locus; F_{IS} – average Wright's inbreeding coefficient.

Pairwise F_{ST} comparisons between populations were calculated in ARLEQUIN from the set of 6006 loci that passed filtering criteria. Of those, 372 contained too much missing data and were not included in the analysis (allowed level of missing data: 0.05). Most values were statistically significant ($P < 0.05$), with the exception of several pairwise comparisons which include the Icelandic and SKA populations (Table 2). Highest F_{ST} values were calculated for pairwise comparisons which included the Turkish population (average $F_{ST} = 0.26$), followed by the population from Spain (average $F_{ST} = 0.11$). Comparisons between the Baltic Sea and adjacent regions yielded markedly lower, yet significant F_{ST} values. Comparisons between NOS and SKA, and the BES and IBS regions showed significant F_{ST} values ranging from 0.022 to 0.026. F_{ST} values between adjacent regions like NOS:KAT and KAT:BES were lower, around 0.011.

Table 2. Pairwise F_{ST} values (above the diagonal) and F_{ST} P values. Values in bold are significant.

	Turkey	Spain	Iceland	NOS	SKA	KAT	BES	IBS
Turkey		**0.3484**	**0.2595**	**0.2371**	0.3008	**0.2454**	**0.2337**	**0.2241**
Spain	**0.0273**		0.0923	**0.1019**	0.1201	**0.1004**	**0.1147**	**0.1126**
Iceland	**0.0244**	0.1992		0.0113	-0.0031	0.0060	**0.0236**	**0.0267**
NOS	**0.0049**	**0.0107**	0.0889		0.0120	**0.0115**	**0.0220**	**0.0260**
SKA	0.0645	0.1777	0.5732	0.2813		0.0082	**0.0216**	**0.0257**
KAT	**0.0059**	**0.0215**	0.1709	**0.0186**	0.2178		**0.0109**	0.0127
BES	**0.0000**	**0.0010**	**0.0059**	**0.0000**	**0.0244**	**0.0049**		0.0043
IBS	**0.0000**	**0.0029**	**0.0361**	**0.0039**	**0.0400**	0.0938	0.2832	

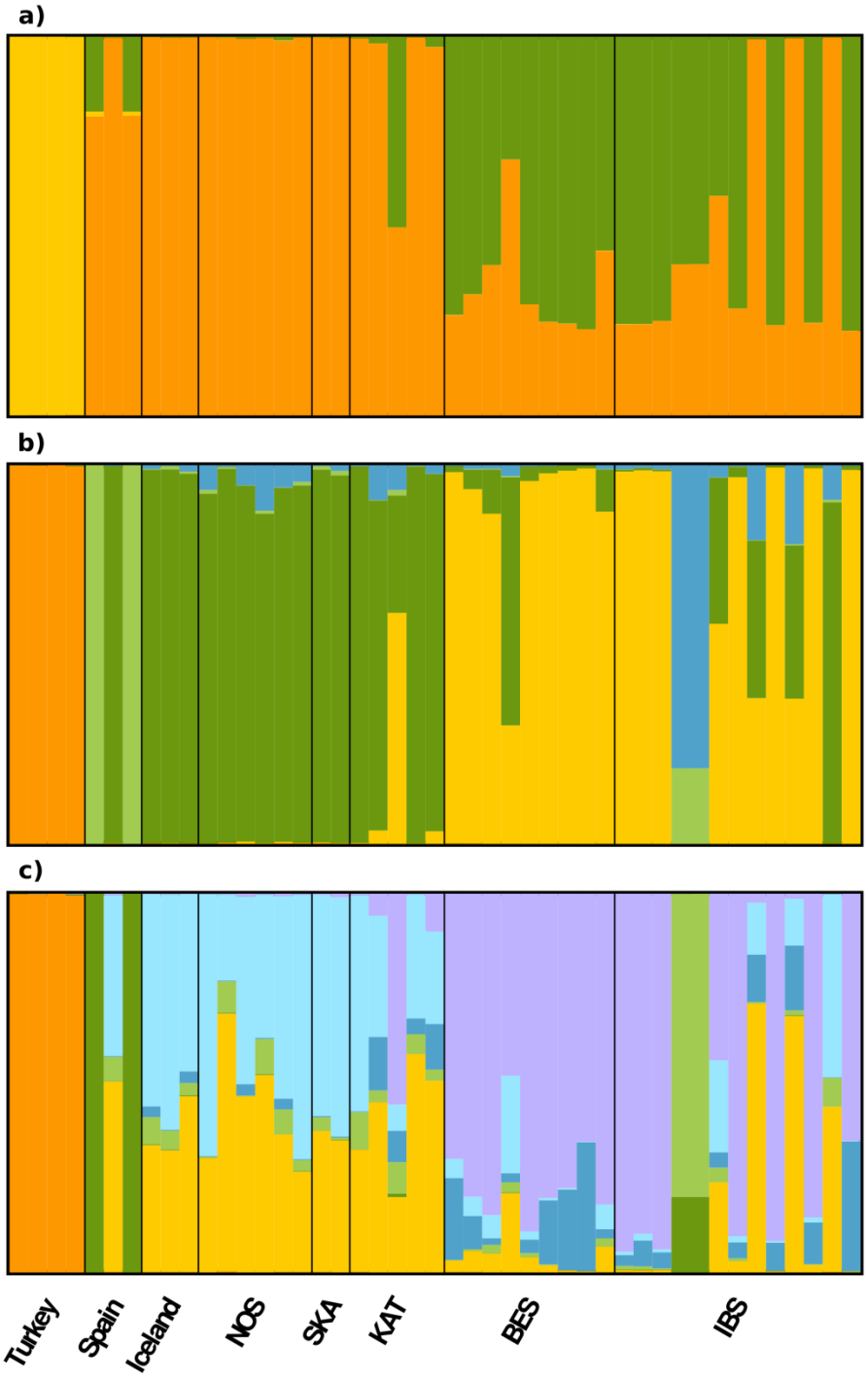

Figure 3. Bayesian plot of group assignment of each individual into three (a), five (b) and seven (c) clusters based on STRUCTURE analyses using 3000 SNPs. Models with K = 3 (delta$K$ = 2.5), K=5 (delta$K$ = 1.3) and K = 7 (delta$K$ = 2.3) best fit the data using the Evanno method. The results are grouped by region of origin. Each of 45 individuals is represented with a vertical column where the coloration is proportional to the individuals estimated membership coefficient in one of the clusters of genetic similarity.

Table3. Occurrence of mtDNA haplotypes among specimens of SNP clusters B and C. The haplotype distribution among the two SNP clusters is significantly different (X^2=9.904, p=0.019).

	Mitochondrial haplotype			
	PHO 1	PHO 4	PHO 7	other haplotype
SNP cluster B	7	3	2	2
SNP cluster C	4	0	11	3
haplotype specific comparison	p=0.366	p=0.083	p=0.013	p=0.655

Harbor porpoise population structure

To investigate harbor porpoise populations structure, we randomly selected 3000 loci for analyses. Because loci in tight linkage, such as those originating from a single RAD site, should be avoided in STRUCTURE analyses (Pritchard *et al*, 2000), only one SNP was chosen from each selected RAD-tag locus. By using the deltaK approach (Evanno *et al*, 2005), we found that models with K = 3 (delta*K* = 2.5), K=5 (delta*K* = 1.3) and K = 7 (delta*K* = 2.3) best fit the data. The plot shows a pattern where at the highest level of structure, the Turkish populations is clearly separated from the North Sea, Baltic Sea and the adjacent populations (Fig. 3a-c). Similarly, two individuals from the Spanish population are clustered together. The third Spanish individual shows a clustering pattern that is found in the Icelandic, NOS and SKA populations. There is a change in cluster representation in the transition between SKA and KAT, and further in the BES populations. The pattern then changes again with transition into the Inner Baltic Sea, with several individuals showing unique clustering. The two specimens forming the Inner Baltic cluster C2 in the k=5 (blue) and k=7 (light green) analysis are two females by-caught east of Sweden in May.

SNP clusters identified by STRUCTURE significantly differed in their mitochondrial haplotype composition (Table 3, see Fig. 3 and Appendix for assignments). There was a significant over-representation of haplotype PHO7 in SNP cluster C and a tendency towards over-representation of PHO4 in SNP cluster B.

DISCUSSION

Our study of population differentiation using the RAD-seq genotyping-by-sequencing successfully reproduced the harbor porpoise population differentiation inferred from 15 microsatellite and mitochondrial control region sequence data (Wiemann *et al*, 2009), including the correlation between nuclear DNA clustering (microsatellites/SNPs) and certain haplotypes of the genetically unlinked mtDNA. Furthermore we achieved a comparatively more precise population assignment of individuals to specific clusters with a much smaller sample set (45 individuals compared to the microsatellite data from 305 samples used in the study mentioned above). Expected clustering was achieved with SNP data from just two individuals (e.g. samples from Turkey, Spain, or SKA). Using the RAD-seq method we were able to obtain a higher resolution with substantially fewer individuals because the method provides a genome-wide sampling of loci that is much denser than with microsatellites (Gärke *et al*, 2012; Haasl & Payseur, 2011; DeFaveri *et al*, 2013; Helyar *et al*, 2011).

Our results show the isolation of Turkish (designated as the subspecies *P. phocoena relicta*), as well as Galician population (the latter however with some affinity to North Sea/North Atlantic). This is in accordance with studies that report limited gene flow in to and out of the porpoise population from Iberian waters, and on a larger scale for porpoises from the Black Sea (Fontaine *et al*, 2007 and references therein). Conversely, the samples from Iceland showed the same assignment pattern as those from NOS and SKA, indicating that there is not much genetic differentiation between porpoises from North Sea and Central North Atlantic. This finding is partly in accordance with Fontaine *et al* (2007) who reported significant isolation-by-distance for the central and eastern North Atlantic population. A much clearer differentiation can be seen between the Icelandic/NOS/SKA clustering, and the clustering in the BES/IBS regions. The Kattegat region appears as a transition zone, with individuals showing clustering patterns characteristic of both adjacent regions. As has been suggested by Fontaine *et al* (2007), barriers separating oceanic basins can cause profound population structure on a small geographic scale. Such might be the case with the Kattegat, Belt Sea and Inner Baltic sea basins separated by shallower underwater ridges (Wiemann *et al*, 2009). Wiemann *et al* (2009) reported that the most striking characteristic of the BES region was the presence of a distinct mtDNA haplotype (PHO7) which rarely appeared in the NOS/SKA region and was absent everywhere else,

indicating a split between the North Sea and the Baltic porpoises. Another split was further suggested between the BES and IBS populations which is to some extent also seen in our data. Furthermore, there appears to be differentiation among individuals within the Inner Baltic Sea with several individuals from the IBS showing a pattern similar to porpoises from the Icelandic/NOS/SKA populations. There are two individuals (females by-caught in May east of Sweden) assigned to a cluster specific to IBS, while others are assigned to Icelandic/NOS/SKA or BES. A scenario that could explain such a clustering pattern would be some seasonal migration between IBS and adjacent areas, but the data do not rule out the possibility of a relict IBS population.

Another important issue to consider is the quality of genomic DNA, as RAD-seq can be limiting in this respect – it requires at least 1 µg of high-quality genomic DNA per sample (Hohenlohe *et al*, 2011). Specifically with regard to harbor porpoise tissue samples it is difficult to acquire or expect 'fresh' samples, since most are collected from stranded individuals and a smaller number are from by-catches (Wiemann et al, 2009; Wright *et al*, 2013). Particularly in the case of strandings, tissue is collected from animals in various stages of decomposition with concomitant decreases in DNA quality. As we have shown in this study, samples of very low DNA quality typically yield a small number of unique RAD-tag loci. If stringency filters are applied, such that all loci for downstream analyses must be present in all samples, then we can expect a smaller data set as output. It is therefore critical to consider what level of genomic DNA degradation is acceptable for a sample to be sequenced given an expected RAD-tag output and the highest possible number of samples.

In summary, this pilot study demonstrates the feasibility of SNP analysis on opportunistically sampled cetacean samples for population diversity and divergence analysis. The ddRAD-seq method delivered around 6000 SNP loci from 49 specimens in a single Illumina lane. Clearly, this approach should be applied to a larger sample set, such that specimens can be stratified by sex and season. Provided a meaningful and sufficiently large set of samples, RAD-tag genotyping has the potential to analyze population differentiation with an unprecedented number of loci, which should yield high resolution power and precision in parameter estimation and population delimitation.

Acknowledgements

Financial support is acknowledged from the University of Potsdam. We thank the Unit of Evolutionary Adaptive Genomics at the University of Potsdam (Prof. Michael Hofreiter) for access to the Agilent Tape Station.

REFERENCES

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA & Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3:** e3376

Catchen J, Hohenlohe PA, Bassham S, Amores A & Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22:** 3124–40

Catchen JM, Amores A, Hohenlohe P, Cresko W & Postlethwait JH (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda).* **1:** 171–82

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM & Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12:** 499–510 Available at: http://dx.doi.org/10.1038/nrg3012 [Accessed March 19, 2014]

DeFaveri J, Viitaniemi H, Leder E & Merilä J (2013) Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. *Mol. Ecol. Resour.* **13:** 377–92

Earl DA & VonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4:** 359–361

Etter PD, Preston JL, Bassham S, Cresko WA & Johnson EA (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS One* **6:** e18561

Evanno G, Regnaut S & Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14:** 2611–20

Excoffier L & Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10:** 564–7

Falush D, Stephens M & Pritchard JK (2003) Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* **164:** 1567–1587

Fontaine MC, Baird SJE, Piry S, Ray N, Tolley KA, Duke S, Birkun A, Ferreira M, Jauniaux T, Llavona A, Oztürk B, A Oztürk A, Ridoux V, Rogan E, Sequeira M, Siebert U, Vikingsson GA, Bouquegneau J-M & Michaux JR (2007) Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters. *BMC Biol.* **5:** 30

Gärke C, Ytournel F, Bed'hom B, Gut I, Lathrop M, Weigend S & Simianer H (2012) Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Anim. Genet.* **43:** 419–28

Haasl RJ & Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity (Edinb).* **106:** 158–71

Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, Cariani A, Maes GE, Diopere E, Carvalho GR & Nielsen EE (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol. Ecol. Resour.* **11 Suppl 1:** 123–36

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW & Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Mol. Ecol. Resour.* **11 Suppl 1:** 117–22

Hubisz MJ, Falush D, Stephens M & Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9:** 1322–32

Jakobsson M & Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23:** 1801–6

Peterson BK, Weber JN, Kay EH, Fisher HS & Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7:** e37135

Pritchard JK, Stephens M & Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155:** 945–959

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M & Belzile F (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* **8:** e54603

Wiemann A, Andersen LW, Berggren P, Siebert U, Benke H, Teilmann J, Lockyer C, Pawliczka I, Skóra K, Roos A, Lyrholm T, Paulus KB, Ketmaier V & Tiedemann R (2009) Mitochondrial Control Region and microsatellite analyses on harbour porpoise (Phocoena phocoena) unravel population differentiation in the Baltic Sea and adjacent waters. *Conserv. Genet.* **11:** 195–211

Wright AJ, Maar M, Mohn C, Nabe-Nielsen J, Siebert U, Jensen LF, Baagøe HJ & Teilmann J (2013) Possible causes of a harbour porpoise mass stranding in Danish waters in 2005. *PLoS One* **8:** e55553

Appendix: Samples in order of the STRUCTURE plot (Figure 3). For the different k values, the STRUCTURE cluster with the highest relative assignment is stated (k=3: A=yellow, B=orange, C=dark green; k=5: A=orange, B1=light green, B2=dark green, C1=yellow, C2=dark blue; k=7: A=orange, B1=dark green, B21=yellow, B22=light blue, C11=violet, C12=dark blue, C2= light green; no specimen was assigned to C12). mt data are from Wiemann et al. 2009).

Structure plot ID	Area	Month	k=3	k=5	k=7	mtDNA haplotype	Sex
1	TUR	/	A	A	A	/	f
2	TUR	/	A	A	A	/	f
3	TUR	/	A	A	A	/	f
4	TUR	/	A	A	A	/	f
5	SP	Feb	B	B1	B1	/	m
6	SP	Feb	B	B2	B21	/	m
7	SP	Feb	B	B1	B1	/	m
8	IS	Jan	B	B2	B22	/	/
9	IS	Apr	B	B2	B22	/	m
10	IS	Jun	B	B2	B22	/	m
11	NOS	Sep	B	B2	B22	1	f
12	NOS	Jun	B	B2	B21	4	f
13	NOS	May	B	B2	B22	1	f
14	NOS	May	B	B2	B21	/	f
15	NOS	May	B	B2	B22	/	f
16	NOS	Apr	B	B2	B22	4	m
17	SKA	Aug	B	B2	B22	1	f
18	SKA	Feb	B	B2	B22	1	m
19	KAT1	Jul	B	B2	B22	7	f
20	KAT2	May	B	B2	B21	27	f
21	KAT2	Jun	BC	C1	C11	7	f
22	KAT2	Jul	B	B2	B21	1	f
23	KAT2	Feb	B	B2	B21	1	m
24	BES1	Aug	C	C1	C11	7	f
25	BES1	Sep	C	C1	C11	7	f
26	BES1	Apr	C	C1	C11	11	f
27	BES1	Nov	B	B2	C11	1	f
28	BES2	Sep	C	C1	C11	7	f
29	BES2	Sep	C	C1	C11	7	f
30	BES2	Aug	C	C1	C11	1	f
31	BES2	Aug	C	C1	C11	14	f
32	BES2	Aug	C	C1	C11	/	f
33	IBS1	Sep	C	C1	C11	1	f
34	IBS1	Aug	C	C1	C11	7	f
35	IBS1	Jul	C	C1	C11	7	f
36	IBS2	May	C	C2	C2	7	f

37	IBS2	May	C	C2	C2	7	f
38	IBS2	Aug	B	C1	C11	1	m
39	IBS3	Nov	C	C1	C11	14	f
40	IBS3	Aug	B	B2	B21	7	m
41	IBS4	Feb	C	C1	C11	7	f
42	IBS4	Jun	B	B2	B21	27	f
43	IBS4	Jul	C	C1	C11	1	m
44	IBS5	Okt	B	B2	B22	4	f
45	IBS5	Jan	C	C1	C11	7	m